

APPENDIX A
FOR
UNITED STATES LETTERS PATENT

TITLE: STRUCTURAL HEALTH MONITORING
APPLICANT: WIESLAW J. STASZEWSKI

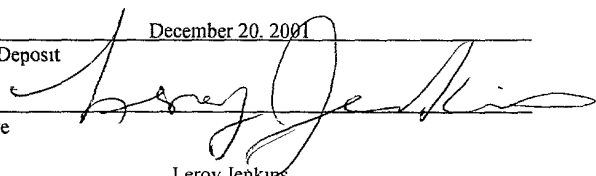
CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EL298428644US

I hereby certify under 37 CFR §1.10 that this correspondence is being deposited with the United States Postal Service as Express Mail Post Office to Addressee with sufficient postage on the date indicated below and is addressed to the Commissioner for Patents, Washington, D.C. 20231.

Date of Deposit December 20, 2001

Signature



Leroy Jenkins

Typed or Printed Name of Person Signing Certificate

Appendix A.

[illegible]

A Survey of Signal Demodulation Algorithms for Fault Detection in Machinery and Structures

S. Patsias and W. J. Staszewski

Dynamics Research Group,

Department of Mechanical Engineering, Sheffield University, United Kingdom.

Mappin Street, Sheffield S1 3JD

Keywords:

Envelope function, instantaneous frequency, machinery diagnostics, structural fault detection

ABSTRACT

This paper brings together a number of algorithms for the estimation of the envelope and instantaneous frequency functions suitable for machinery diagnostics and structural health monitoring. This includes the classical approach based on Rice's frequency, the most commonly used method based on the Hilbert transform and recent developments related to the Wigner-Ville distribution and wavelet analysis. The study includes two application examples from fault detection in gearboxes and identification systems. The paper shows that the envelope and instantaneous frequency functions are an important link between classical Fourier approach based methods and time-variant analysis.

1. INTRODUCTION

The viewpoint of this paper is that there exist three major areas within machinery diagnostics and structural health monitoring where the instantaneous characteristics, namely the envelope and instantaneous frequency, are used. These are: damage detection in rotating machinery, identification of nonlinear systems and deterministic approach to stationarity.

Machines and structures in operation generate forces and motions that produce different types of vibration. Any fault in machine operation or structural damage can be considered as an additional excitation which results in a change of vibration response. Often the overall process of vibration response exhibit modulations, intermodulations or nonlinear behaviour due to operational nature or fault. Identification of modulation sources can provide valuable information of many faults in rotating machinery, particularly in gearboxes and ball-bearings. Modulations in gearboxes are mainly caused by the tooth meshing process, due to varying stiffness, or by the eccentricity and local tooth faults¹. This is exhibited by modulation sidebands in vibration spectra.

Intermodulations results from sum and difference frequencies of low frequency harmonics of the shaft speed, harmonics of the tooth meshing frequency and modulation sidebands. Modulations in ball-bearings are mainly caused by local surface defects which produce impulses often leading to resonances. For advanced defects, simple parameters such as kurtosis or crest factor^{2,3} are sufficient to detect damage. Often modulation and intermodulations which result from damage can be identified by means of ZOOM-spectrum and cepstrum^{4,5}. For complex machines the signal demodulation procedure, or in other words, the analysis of the envelope and instantaneous frequency functions, is very effective to identify local tooth damage⁶⁻¹⁰ and local defects in ball-bearings^{11-14,9-10}. The amplitude demodulation procedure is often called envelope detection or, when used in ball-bearings, the high-frequency resonance technique.

The second important area where the envelope and instantaneous frequency functions are often used is identification of nonlinear systems. Since nonlinearities may result from damage in a system, the distinction between linear and nonlinear behaviour is an

important problem in damage identification. Many different procedures for the identification of nonlinear systems from vibration test data have been developed in structural dynamics. It is well known that many types of nonlinearities cause a varying nature for the restoring forces and natural frequencies of the system. This varying nature of vibration can be studied using the envelope and instantaneous frequency functions which lead to nonlinear back-bone and damping curves.

The overall vibration produced by machinery and structures is either stationary or nonstationary. The analysis of the envelope and instantaneous frequency functions often forms the deterministic test for stationarity in vibration analysis¹⁵. In this approach, the process is said to be stationary if it consists of components such that their instantaneous behaviour, as described by the envelope and instantaneous frequency, does not depend on time. Conversely, the process is non-stationary if its instantaneous characteristics depend on time. The instantaneous characteristics are used in machinery diagnostics to detect local faults. This fault detection method is in fact a test of stationarity and any departure of instantaneous characteristics from time independent functions is considered as a symptom of a fault in the system.

Recent developments in signal processing show that the envelope and instantaneous frequency functions are an important link between classical Fourier approach based methods and time-variant analysis, as shown in ¹⁵. There exist a number of procedures, developed over the last twenty years, which can be used to estimate the envelope and instantaneous frequency functions. It appears that the Hilbert transform based procedure¹⁶ is the most common approach to the problem. Recent developments include the Wigner-Ville distribution¹⁷ and the wavelet transform¹⁸. However, there exist a number of different algorithms which can be used for single- and multi-component signals (frequencies).

The aim of this paper is to bring together the algorithms of signal demodulation. It is hoped that this form of presentation will help to implement the most suitable approach for a

given vibration problem. However, the paper does not intend to survey various application examples.

The structure of the paper is as follows. Section 2 shows the complexity of modulation and intermodulation processes which can be found in rotating machinery. The algorithms of the amplitude and frequency/phase demodulation are given in Sections 3 and 4, respectively. This is followed by Section 5 with examples related to fault detection in gearboxes and identification of nonlinear systems. Finally, the paper is concluded in Section 6.

2. MODULATION AND INTERMODULATION PROCESSES IN MACHINE DIAGNOSTICS

In machinery, modulation is the variation in the value of some parameter which characterizes a periodic oscillation. This process is an effect of either design features of machines (e.g. crankshafts) or more often a result of a fault. Thus the identification of modulation sources can provide valuable information about possible faults in machinery. Amplitude modulation processes, in general, are caused by impacts and impulses in the structure, generated by different phenomena, for example, cracks or missing teeth in gears, point defects in ball-bearings, blade cracks in fans or any other local faults. The impulses are repeated periodically for each revolution of the wheel, shaft, etc. The frequency modulation processes are an effect of parametrical excitation in the structure (e.g. fluctuation in tooth loading in gearboxes, varying stiffness of teeth in gears, varying cross-section of shafts, etc.). More details can be found in^{1,19-25}. In what follows some spectral properties of modulated processes will be relevant to machinery vibration as summed.

The vibration signal generated by a machine can be analytically described as a complex wave

$$x(t) = \sum_{r=1}^R [1 + a_r(t)] e^{j[\omega_r t + \phi_r(t)]} + n(t)$$

(1)

where $a_r(t)$ is the slow-varying amplitude modulation process, $\phi_r(t)$ is the fast varying phase modulation process, f_{or} are the carrier frequencies ($\omega_{or} = 2\pi f_{or} = \text{const}$) and $n(t)$ is the noise. Amplitude and phase modulation processes can be represented by,

$$a_r(t) = \sum_{k=1}^{n_r} M_{ir} \cos \Omega_{ir} t$$

(2)

$$\phi_r(t) = \sum_{k=1}^{m_r} \beta_{kr} \cos \omega_{kr} t$$

(3)

where M_{ir} are the coefficients of amplitude modulation intensity and, β_{kr} are phase deviations. Substituting (2) and (3) into Eq. (1), it follows that

$$x(t) = \sum_{r=1}^R \left[1 + \sum_{i=1}^{n_r} M_{ir} \cos \Omega_{ir} t \right] e^{j\omega_{or} t + \sum_{k=1}^{m_r} j\beta_{kr} \cos \omega_{kr} t} + n(t)$$

(4)

Eq. (2) can be replaced by,

$$a_r(t) = \frac{1}{2} \sum_{i=1}^{n_r} M_{ir} (e^{j\Omega_{ir} t} + e^{-j\Omega_{ir} t})$$

(5)

The well known Fourier-series expansion,

$$e^{j\beta_{kr} \cos \omega_{kr} t} = \sum_{n=-\infty}^{+\infty} j^n J_n(\beta_{kr}) e^{jn\omega_{kr} t}$$

(6)

where $J_n(\beta_{kr})$ is the Bessel function of the first kind, together with Eqs. (4) and (5) yields,

$$x(t) = \sum_{r=1}^R \left\{ \left[1 + \frac{1}{2} \sum_{i=1}^{n_r} M_{ir} (e^{j\Omega_{ir}t} + e^{-j\Omega_{ir}t}) \right] e^{j\omega_{or}t} \left[\sum_{n=-\infty}^{+\infty} j^n J_n(\beta_{1r}) e^{jn\omega_{1r}t} \right] \right. \\ \left. \left[\sum_{n=-\infty}^{+\infty} j^n J_n(\beta_{2r}) e^{jn\omega_{2r}t} \right] \dots \left[\sum_{n=-\infty}^{+\infty} j^n J_n(\beta_{m_r,r}) e^{jn\omega_{mr}t} \right] \right\} + n(t)$$

(7)

$J_n(\beta_{kr})$ in Eq. (6) is the Bessel function of the first kind. Eq. (7) can be written as,

$$x(t) = \sum_{r=1}^R \left\{ \left[1 + \frac{1}{2} \sum_{i=1}^{n_r} M_{ir} (e^{j\Omega_{ir}t} + e^{-j\Omega_{ir}t}) \right] \left[\prod_{k=1}^{m_r} \sum_{n=-\infty}^{+\infty} j^n J_n(\beta_{kr}) e^{jn\omega_{kr}t} \right] e^{j\omega_{or}t} \right\} + n(t)$$

(8)

or finally,

$$x(t) = \sum_{r=1}^R \left\{ \left[1 + \frac{1}{2} \sum_{i=1}^{n_r} M_{ir} (e^{j\Omega_{ir}t} + e^{-j\Omega_{ir}t}) \right] \left[\prod_{k=1}^{m_r} \sum_{n=-\infty}^{+\infty} j^n J_n(\beta_{kr}) e^{j(\omega_{or} + n\omega_{1r} + n\omega_{2r} + \dots + n\omega_{mr})t} \right] \right\} + n(t)$$

(9)

Analysing Eq. (9) one can notice that in the frequency domain the vibration signal $x(t)$ is represented by R families of spectral components. Each family yields five types of side-frequency terms:

$A_1 e^{j\omega_{or}t}$ - carrier frequency

$A_2 e^{j(\omega_{or} \pm \Omega_{ir})t}$ - side frequencies due to amplitude modulating waves

$A_3 e^{j(\omega_{or} \pm n\omega_{kr})t}$ - side frequencies due to phase modulating wave

$A_4 e^{j(\omega_{or} \pm n\omega_{1r} \pm n\omega_{2r} \pm \dots \pm n\omega_{kr})t}$ - beat side frequencies due to phase modulation

$A_5 e^{j(\omega_{or} \pm \Omega_{1r} \pm \Omega_{2r} \pm \dots \pm \Omega_{ir} \pm n\omega_{1r} \pm n\omega_{2r} \pm \dots \pm n\omega_{kr})t}$ - beat side frequencies due to amplitude and phase modulation,

where A_i are the Bessel function components according to Eq. (9). Fig. 1 shows an example of the amplitude and phase modulated signal, which includes two families of spectral components, and its spectrum.

In the case of the pure amplitude modulation process ($\omega_{kr} = 0$) the total energy of the signal will increase. The energy of the carrier remains unchanged and the additional energy which comes from the modulation wave will be represented in sidebands. This process yields only one pair of sidebands: sum and difference components. In comparison the total energy of the phase (frequency) modulated wave remains constant. To achieve this the energy of the carrier will decrease and the difference will be distributed to its infinite number of sidebands. In addition phase (frequency) modulation can produce beat components, which are not characteristic for the amplitude modulation. In practice, for example, the carrier frequency in the gearbox deals with the meshing frequency. Thus one should remember that greater advancement of a fault does not need to cause the increase of the level of the meshing frequency. This means that the level of the meshing frequency might not be a good indication of the advancement of a fault. Additionally, if one considers the phase angle between the

amplitude and phase modulation waves i.e. $a_r(t) = \sum_{i=1}^{n_r} M_{ir} \cos(\Omega_{ir}t + \Omega_{ir0})$, thus the

asymmetric distribution of sidebands will be observed. These brief considerations show that the spectrum of the vibration signal generated by a complex machine might be very

complicated. Fig. 1 shows an example of amplitude and phase modulated signal and its spectrum.

3. AMPLITUDE DEMODULATION ALGORITHMS

Amplitude demodulation is the process of extracting slow variation of the amplitude from the modulated signal. This operation falls into two broad categories of synchronous detection and envelope detection²⁶. Both categories are summarised below. Additionally, the algorithm involving Discrete Fourier Analysis (DFA)²⁷ is also described.

3.1. Synchronous Detection

Synchronous detection uses the operation of frequency conversion. The concept is outlined in Fig. 2. Here the oscillator is assumed to be exactly synchronised with the carrier frequency.

Assume an amplitude modulated signal, $x(t) = a(t)\cos\omega_o t$ where ω_o is the carrier frequency. Multiplying both sides by $2\cos\omega_o t$ one obtains

$$2x(t)\cos\omega_o t = a(t) + a(t)\cos 2\omega_o t$$

(10)

The first term on the right side is the one we want, since it is an amplitude modulation signal. The second term can be removed by low-pass filtering. This procedure requires the carrier frequency determination in order to translate the signal spectrum to a new carrier frequency equal to zero. Nevertheless some degree of asynchronism must be expected,

since the carrier frequency cannot always be exactly determined (for example in the case of varying rotational speed of the gearbox). This means that instead of multiplying by $2\cos\omega_o t$ one may use the value $2\cos(\omega_o + \omega'_o)t$ which results in,

$$2x(t)\cos(\omega_o + \omega'_o)t = a(t)\cos\omega'_o t + a(t)\cos(2\omega_o + \omega'_o)t$$

(11)

Here the second term can still be removed but the harmonic relationships in the modulation signal $a(t)$ are destroyed. Therefore instead of $a(t)$ one receives $a(t)\cos\omega'_o t$, which in the frequency domain, according to the modulation theorem²⁸, gives us the shifted spectrum $A\left(f + \frac{\omega'_o}{2\pi}\right)$, without any possibility of correction. The second difficulty with this procedure is related to the digital filtering operation which reduces the efficiency of the algorithm.

An alternative implementation of this method is presented in section 4.2, where the frequency modulation signal $\phi(t)$ is assumed not to be equal to zero, which allows one to compute either $a(t)$ or $\phi(t)$.

3.2. Envelope Detection

There exist a number of envelope definitions. All of them come from random vibration or communication systems theory. Physically, for the narrow-banded process $x(t)$, the envelope is a smooth curve joining the peaks of $x(t)$. A few envelope definitions and their statistical properties have been described by Langley²⁹. In what follows a few possible implementations of envelope detection methods are briefly discussed.

3.2.1. Local maxima distribution

The envelope as a local maxima distribution was defined by Crandall³⁰⁻³¹ as being a smooth gradial curve joining the peaks and the expected time spent by the envelope between the level a and $a+da$ is just the number of those peaks, whose magnitude lies between a and $a+da$ multiplied by the expected period for cycles of amplitude a . In practice the method leads to running averages of signal local maxima. If the modulated process is written as $x(t)=a(t)\cos[\omega_0 t+\phi(t)]$, the running averages can be carried out as³²

$$e(t) = E_{\tau} \{ \text{Max}[x(t)] \} = \frac{1}{\tau} \int_{\tau}^{\tau+t} \text{Max}\{a(t)\cos[\omega_0 t + \phi(t)]\} d\tau \quad (12)$$

where E_{τ} is a symbol of averaging. If the phase process is constant ($\phi(t)=\text{const}$), for the optimal averaging time $\tau \gg \frac{2\pi}{\omega_0}$, Eq. (12) can be replaced by

$$e(t) = \frac{1}{\tau} \int_{\tau}^{\tau+t} a(t) dt = a(t) \quad (13)$$

This definition implies the bandwidth restriction. The efficiency of the method depends on the choice of the optimal averaging time. It also increases if the process is more narrow-banded. The local maxima finding procedure can be done by the approximate construction of tangents structure. Unfortunately the envelope is then given as a series of finite number samples, which makes exact tangent points calculations impossible. Additionally, the integration procedure requires a few samples in a period to reach a good accuracy³³.

3.2.2. Energy based distribution

The energy based distribution envelope proposed by Crandall³⁰⁻³¹ is related to stochastic processes. If the stochastic equation of motion is assumed to be of the form³⁰

$$\ddot{x} + \beta A(x, \dot{x}) + G(x) = F(t)$$

(14)

where x is a displacement response process, A is assumed to be an odd function with respect to \dot{x} , β is a small parameter, $F(t)$ is the white noise excitation, the envelope function $a(t)$ of the random process $x(t)$ is given implicitly by²⁰

$$a(t) = \frac{1}{2} \dot{x}^2 + V(x)$$

(15)

where $V(x) = \int_0^x G(\zeta) d\zeta$. The right-hand side of the above equation is the sum of the kinetic and potential energy per unit mass. In the linear case one can set $\beta = 2\zeta\omega_c$, $F(x, \dot{x}) = \dot{x}$, $G(x) = \omega_c^2 x$ and Eq. (15) can be replaced by

$$a(t) = \sqrt{x^2(t) + \left(\frac{\dot{x}(t)}{\omega_c}\right)^2}$$

(16)

where ω_c is some constant frequency. Langley²⁸ has suggested to use ω_c as the mean zero crossing frequency given by $\omega_c = \sqrt{\frac{m_2}{m_0}}$, where m_n is the n -th spectral moment of the single sided spectrum of $x(t)$ defined as

$$m_n = \int_0^\infty \omega^n S_{xx}(\omega) d\omega$$

(17)

This definition of the envelope does not imply any bandwidth restriction. It is easy to notice that such an envelope follows the peaks of $x(t)$, since the envelope and the process coincide at a peak, where $\dot{x}(t) = 0$. The implementation of the method can be realized following Eq. (16). The effect of the algorithm is shown in Fig. 3. This requires signal

differentiation, which is not a simple task³³. The differentiation can be avoided by measuring acceleration or velocity where possible and integrating to obtain velocity or displacement. The accuracy still requires sampling at over a few times the highest frequency of interest³³.

Langley²⁹ has shown that the mean rate at which the envelope crosses a given level with positive slope depends on the 4-th spectral moment m_4 . He indicated that some random processes, such as for instance the response of a linear system to white noise, have a theoretically infinite value of m_4 . This predicts that the envelope crossing rate is infinite. In practice signals generated by machines have a finite value of m_4 , if not, it is still possible to filter the response spectrum at a frequency which yields a finite value of m_4 .

3.2.3. Rice's envelope

The so-called classical definition of the envelope has been given by Rice³⁴. Writing the modulated signal in the form

$$x(t) = \sum_n a_n(t) \cos[\omega_n t + \phi_n(t)]$$

(18)

and selecting a frequency f_m called the "midband frequency" Eq. (18) can be replaced by

$$x(t) = \sum_n a_n(t) \cos[(\omega_n - f_m)t + \phi_n(t) + f_m t] = I_C \cos f_m t - I_S \sin f_m t$$

(19)

where,

$$I_C = \sum_n c_n \cos[(\omega_n - f_m)t + \phi_n(t)]$$

(20)

$$I_S = \sum_n c_n \sin[(\omega_n - f_m)t + \phi_n(t)]$$

The expression

$$a(t) = \sqrt{I_C^2 + I_S^2}$$

(21)

is termed by Rice the envelope of $x(t)$ referred to frequency f_m . It has been shown by Dugundji³⁵ that this envelope is completely independent of the midband frequency f_m . The explicit calculation of Eq. (21) is rather impossible. Thus the Rice's envelope is presented only as a theoretical case.

3.2.4. Hilbert transform approach

The modulated signal $x(t)$ can be replaced on the basis of the analytic signal $x_H(t)$ ³⁶

$$x_A(t) = x(t) + jx_H(t)$$

(22)

where $x_H(t)$ is the Hilbert transform of $x(t)$ being²⁸

$$H[x(t)] = x(t) = \frac{1}{\pi} \int_{-\infty}^{+\infty} x(\tau) \frac{1}{t-\tau} d\tau$$

(23)

Thus the envelope suggested by Dugundji³⁵ is possible

$$a(t) = \sqrt{x^2(t) + x_H^2(t)}$$

(24)

It has been shown^{29,35} that this envelope and Rice's envelope are equivalent and the proof of equivalence does not depend upon $x(t)$ being Gaussian. It is easy to notice that for $x(t)$ being harmonic ($x(t) = A \cos \omega_o t$) the energy based envelope (16) and the envelope defined by (24) are also equivalent. However, the central carrier frequency of Dugundji's envelope is independent of the carrier frequency.

Langley²⁹ has assessed when the envelope given by Dugundji (24) will follow the peaks of $x(t)$. The mean and variance of $x_H(t)$ when $\dot{x}(t)$ has a specific value is given by²⁹

$$E\left[\frac{x_H}{\dot{x}}\right] = -\left(\frac{m_1}{m_2}\right) \dot{x}$$

(25)

$$\text{var}\left[\frac{x_H}{\dot{x}}\right] = m_o q^2$$

(26)

where q is a parameter which measure the extend to which $x(t)$ is narrow-banded

$$q^2 = 1 - \frac{m_1^2}{m_o m_2}$$

(27)

Thus $a(t)$ given by (24) will follow the peaks of $x(t)$ ($\dot{x}=0$) if q is small ($x_H=0$) which requires $x(t)$ to be the narrow banded process. It has been shown that the mean rate at which the envelope (24) crosses a given level with positive slope is always finite, which does not take place in the energy based envelope.

The signal processing implementation of the envelope proposed by Dugundji³⁵ can be realised according to Eq. (24). The Hilbert transform definition given by Eq. (23) can be written in the convolution form²⁸

$$H[x(t)] = x_H(t) = \frac{1}{\pi} x(t) * \frac{1}{t}$$

(28)

Taking into account the signum function,

$$\text{sgn } t = \begin{cases} -1 & t < 0 \\ +1 & t > 0 \end{cases}$$

(29)

and the Fourier transform²⁸,

$$F[x(t)] = X(f) = \int_{-\infty}^{+\infty} x(t) e^{-2\pi j f t} dt$$

(30)

the following is obtained

$$F[\text{sgn } t] = \frac{1}{j\pi f}$$

(31)

$$F\left[-\frac{1}{j\pi f}\right]$$

(32)

Thus Eqs. (28), (31) and (32) yield

$$F[x_H(t)] = X_T(f) = X(f)(-j \text{sgn } f)$$

(33)

where X_T is the signal $X(f)$ with shifted phase by $\frac{\pi}{2}$ for the negative frequency components and $-\frac{\pi}{2}$ for the positive frequency components. It means, that the Hilbert transform $x_H(t)$ of the vibration signal $x(t)$ can be easily obtained by calculating the complex spectrum of the signal $X(f)$, shifting the phase $(X_T(f))$ according to Eq. (33) and employing the inverse Fast Fourier Transform (FFT). The computation procedure can be realized also in another version. Going back to Eq. (22) and using the Fourier transform (30) one can find

$$X(f) = X(f) + jX_T(f) = X(f) + \operatorname{sgn} f X(f) = \begin{cases} 0 & f < 0 \\ X(f) & f = 0 \\ 2X(f) & f > 0 \end{cases}$$

(34)

This equation represents the spectrum of the analytic signal and can be very easy computed using the spectrum of the signal $x(t)$. It allows the computation of the envelope function directly from Eq. (22) as a modulus of the analytic signal.

The method presented in this section seems to be very effective and is often given as an example of an effective algorithm of the amplitude demodulation^{37,21,38,9}. It has been used in a number of diagnostics applications^{37,21,38,9,32}.

3.2.5. Discrete Fourier analysis (DFA) algorithm

This algorithm has been formulated and tested by Hsueh and Bielawa²⁷. The assumption is that the amplitude and phase are very slowly relative to the carrier frequency, which means that signals are narrow-banded. If they are additionally periodic, they can be represented by a time-variable Fourier series. Thus a typical amplitude modulated signal can be written as follows²⁷

$$x(t) = x_0(t) + x_{mc}(t) \cos \omega_0 t + x_{ms}(t) \sin \omega_0 t + \text{higher harmonic terms} + \text{noise}$$

(35)

where ω_0 is the carrier frequency, x_0, x_{mc}, x_{ms} are the time-variable Fourier coefficients of $x(t)$, representing the envelope functions. According to the discrete time-variable version of the Fourier analysis. These harmonic coefficients are given by²⁷

$$\begin{aligned} x_0(t) &= \frac{1}{T_c} \int_{t_k - T_c}^{t_k} x(t) dt \\ x_{mc}(t) &= \frac{2}{T_c} \int_{t_k - T_c}^{t_k} x(t) \cos \omega_0 t dt \\ x_{ms}(t) &= \frac{2}{T_c} \int_{t_k - T_c}^{t_k} x(t) \sin \omega_0 t dt \end{aligned}$$

(36)

where T_c is the carrier period and t_k is the time instant. Thus in order to detect the amplitudes of the envelope parts of the input signals, it is necessary at first to remove, by means of analogue and/or digital initial filtering, the higher harmonic terms and high frequency noise from Eq. (36). Then the DFA analysis according to Eq. (37) can be performed. It is easy to notice that this algorithm is similar to the synchronous detection method described in section 3.1. Instead of low-pass filtering after frequency conversion, the DFA algorithm has been used. The implementation of this method²⁷ has shown its very good accuracy and timing characteristics. However, a large number of samples in one carrier period is also required because of a numerical integration. This method also needs the carrier frequency determination. Thus some degree of the carrier frequency shift must be expected and the analysis similar to that one presented in section 3.1 can be performed.

4. FREQUENCY DEMODULATION ALGORITHMS

The frequency demodulation procedure is connected with the computation of an instantaneous frequency $f(t)$ of the signal $x(t)$. According to Eq. (1) this operation can be written as,

$$\frac{\beta}{2\pi} f(t) = \frac{1}{2\pi} \frac{d\theta}{dt} - f_0 \quad (37)$$

where $\theta(t)$ is the total instantaneous phase of the signal and f_0 is the carrier frequency. Since this operation requires signal differentiation, in many cases frequency demodulation is replaced by phase the demodulation procedure which gives

$$\beta \phi(t) = \theta(t) - \omega_0 t \quad (38)$$

where $\omega_0 = 2\pi f_0$. Eq. (38) shows that phase demodulation requires removing of the constant rotational term $\omega_0 t$ related to the carrier frequency. Thus the required carrier frequency can be determined. Both approaches are similar from the diagnostic point of view since they give similar information about the fault. In contrast to amplitude demodulation, frequency (phase) demodulation is a nonlinear operation because of the nonlinear relationship between the signal $x(t)$ and its instantaneous frequency (phase). Thus the procedure is much more difficult.

There exist many different algorithms of frequency and phase demodulation. Almost all of them are included in one of the following categories²⁶: (a) FM-to-AM conversion, (b) phase-shift discrimination, (c) zero-crossing detection and (d) frequency feedback. A signal processing implementation of these algorithms are presented below.

4.1. FM-to-AM Conversion

Any device whose output equals the time derivative of the input produces frequency modulation-to-amplitude modulation conversion²⁸. Consider the frequency modulated signal

$$x(t) = A \cos[\omega_0 t + \phi(t)] \quad (39)$$

Then

$$\begin{aligned} \frac{dx(t)}{dt} &= -A \left(\omega_0 + \frac{d\phi(t)}{dt} \right) \sin[\omega_0 t + \phi(t)] \\ &= 2\pi A \left(f_0 + \frac{1}{2\pi} \frac{d\phi(t)}{dt} \right) \cos[\omega_0 t + \phi(t \pm 180^\circ)] \end{aligned} \quad (40)$$

Thus taking into account Eq. (38) one can obtain

$$f(t) = \frac{Env|\dot{x}(t)|}{2\pi A} - f_0 \quad (41)$$

where $Env|\dot{x}(t)|$ is the envelope function of the differentiated signal $\dot{x}(t)$. Input signal is assumed to have some constant amplitude A . In practice an amplitude limiter is necessary at the input to remove any variations. Such a limiter is not easy to implement. The envelope function can be calculated by means of any algorithm presented in section 3. Additionally signal differentiation is required. This can be avoided, by analogy to section 3.2.2, by measuring acceleration and integration to obtain velocity. This method has been applied in machinery diagnostics^{10,32}.

4.2 Synchronous Detection

Synchronous detection used in section 3.1 to compute amplitude modulation signal can be, by analogy, used also in the case of phase modulation. Considering $x(t) = A \cos[\omega_0 t + \phi(t)]$ and multiplying both sides by $2 \cos \omega_0 t$ one obtains

$$2x(t)\cos\omega_0 t = A\cos\phi(t) + A\cos[2\omega_0 t + \phi(t)]$$

(42)

The second term can be, by analogy to section 3.1, removed by low-pass filtering. Thus the instantaneous phase process is easy to extract. The method used simultaneously for amplitude and phase demodulation purpose, is very often referred to as complex demodulation. Thus the modulated signal is multiplied by $2e^{-j\omega_0 t}$, giving

$$2e^{-j\omega_0 t} a(t)\cos[\omega_0 t + \phi(t)] = a(t)e^{j\phi(t)} + a(t)e^{-j[2\omega_0 t + \phi(t)]}$$

(43)

Then the procedure to obtain demodulated signals fall onto the following steps³⁹:

(a) determine carrier frequency from the spectrum and bandwidth $\Delta\omega$ occupied by sidebands, (b) multiply time series by $2e^{-j\omega_0 t}$, (c) apply a digital low-pass filter and (d) compute the amplitude and phase of the complex signal obtained from 3.

More details about this method can be found in⁴⁰. Either in communication systems or in signal processing complex demodulation has received relatively little attention. The main difficulty is related to the carrier frequency determination. The carrier frequency asynchronous destroys harmonic relations in demodulated spectra.

4.3 Hilbert Transform Approach

The analytic signal used in section 3.2.4 to define the envelope function is the basis of this approach. Going back to Eq. (22) one can write

$$x_A(t) = x(t) + jx_H(t) = a(t)e^{j\theta(t)}$$

(44)

where $a(t)$ is the envelope described in section 2.2.4 and $\theta(t)$ is the total instantaneous phase of the signal being,

$$\theta(t) = \arctan \frac{x_H(t)}{x(t)} = \omega_0 t + \phi(t)$$

(45)

Taking into account the method of the digital amplitude demodulation described in section 3.2.4, the phase demodulation can be realized. Both operations can be realized simultaneously (Fig.4). This algorithm is well established and often used in practice^{37,21,38,9}. The method requires removing the carrier frequency part from the total instantaneous phase. This can be done by translating the spectrum components to the negative frequencies, so that the component originally at $\frac{\omega_0}{2\pi}$ will be moved to zero. Since in practice the carrier frequency is determined with some error $\frac{\Delta\omega_0}{2\pi}$ instead of $\phi(t)$ one receives $\phi(t) + \frac{\Delta\omega_0}{2\pi}$. Using the Fourier transform and multiplication theorem of the Fourier transform²⁸, one will receive in the frequency domain the spectrum of the instantaneous phase

$$\Phi'(f) = \Phi(f) + \frac{\Delta\omega_0}{-j4\pi^2} \frac{d\delta(f)}{df}$$

(46)

which gives an unwanted component in the low frequencies. There exist three methods to improve the results. First, the differentiation operation of the total instantaneous phase $\theta(t)$ can be performed which gives us the instantaneous frequency $f(t)$. This operation is not effective at all. The other two ways of removing low-frequency trends are³³ least-squares polynomial trend removal and high-pass filtering. Since the phase modulation processes are low-frequency type, using high-pass filtering, one should be very careful in order not to remove the required data. In practice the polynomial trend removal is preferable³³.

4.4. Zero-crossing Detection

The zero-crossing frequency can be described on the basis of Rice's frequency defined as¹⁹

$$f_R = \sqrt{\frac{\int_{-\infty}^{+\infty} f^2 S_{xx}(f) df}{\int_{-\infty}^{+\infty} S_{xx}(f) df}}$$

(47)

From the physical point of view this value indicates the central frequency on which the whole power is concentrated. For a narrow-banded Gaussian process Rice's frequency is equal to the expected rate of zero-crossings with positive slope, which can be written

$$E[N_+(0)] = \lim_{T \rightarrow \infty} \frac{N}{\Delta \tau}$$

(48)

where N is a number of "positive" zero-crossing values and $\Delta \tau$ is the interval of time. In practice Eq. (48) can be replaced by

$$\omega(\Delta \tau) = \frac{\Delta \phi}{\Delta \tau}$$

(49)

Thus practically, it resolves into calculations of the interval of time for which the change of the phase has a 2π value, which gives us an estimated instantaneous frequency. This method used in VA diagnostics²¹ requires a great number of samples in one carrier period and does not give sufficient accuracy. However, it does not require the carrier frequency determination.

4.5. Wigner-Ville Distribution

The Wigner distribution (WD) can be derived by generalising the relationship between the power spectrum and the autocorrelation function for non-stationary, time-variant processes. The physical interpretation of the generalised power spectrum $F(\tau, f)$ is that it represents the instantaneous power density spectrum. This leads to the WD defined as⁴¹,

$$W_x(\tau, f) = \int_{-\infty}^{+\infty} x^*\left(t - \frac{\tau}{2}\right) x\left(t + \frac{\tau}{2}\right) e^{-i2\pi f \tau} d\tau \quad (50)$$

It can be shown that the first derivative of an arbitrary signal's phase reflects the mean instantaneous frequency. The WVD can also be used to extract the instantaneous frequency:

Let $x(t) = A(t)e^{i\phi(t)}$ where $A(t)$ and $\phi(t)$, the magnitude and phase respectively are real valued functions, the first derivative of the phase is given by⁴¹,

$$\langle f \rangle_x = \frac{\frac{1}{2\pi} \int f WVD_x(t, f) df}{\frac{1}{2\pi} \int WVD_x(t, f) df} = \frac{\frac{1}{2\pi} \int f WVD_x(t, f) df}{|A(t)|^2} = \phi'(t) \quad (51)$$

This simply means that at time t the mean instantaneous frequency of the signal is equal to the mean instantaneous frequency of the WVD. The disadvantage is that at any time instant t there is more than one frequency component present, i.e. the instantaneous frequency is not a single value, signal energy spreads with respect to the mean instantaneous frequencies.

4.6. Wavelet Transform: Zero-crossings Detection

The wavelet transform is used to decompose a signal $x(t)$ into wavelet coefficients $(\mathcal{W}_\psi x)(a, b)$ using the basis of wavelet functions $\psi_{a,b}(t)$. This can be expressed as,

$$(\mathcal{W}_\psi x)(a, b) = \int_{-\infty}^{+\infty} x(t) \psi_{a,b}^*(t) dt \quad (52)$$

where $\psi^*(.)$ is the complex conjugate of $\psi(.)$. There are many functions that can be used as the wavelet basis functions, an example is the Morlet wavelet⁴².

The square of the modulus of the wavelet transform can be interpreted as an energy density over the (a, b) time-scale plane. The energy of a signal is mainly concentrated on the time-scale plane around the so called ridge of the wavelet transform.

It can be shown that^{18,15},

$$\bar{\phi}_{a,b}(t) = \phi_x(t) - \phi_\psi\left(\frac{t-b}{a}\right) \quad (53)$$

where ϕ_x and ϕ_ψ denote the instantaneous phases of the signal and the wavelet transform respectively. The ridge of the wavelet transform is directly related to the instantaneous frequency of the signal. The ridge is defined as,

$$r(b) = \frac{\dot{\phi}_\psi(0)}{\dot{\phi}_x(0)} \quad (54)$$

This is used to obtain the instantaneous frequency directly from the ridge of the wavelet transform. There exist two algorithms based on the amplitude and phase of the

transform⁴³. Application examples include damping estimation procedures⁴⁴⁻⁴⁵ and identification of nonlinear systems⁴⁴.

5. EXAMPLES AND APPLICATIONS

Two simple examples are shown here to illustrate the application of the instantaneous characteristics. The first example shows the power of the Hilbert transform for damage detection in gearboxes. The second example from the area of system identification shows that often wavelet analysis is better when the instantaneous frequency rather than phase is required in calculations.

5.1 Fault Detection in Gearboxes

Fault characteristics in gearboxes appear in the modulated non-stationary form of impacts, which are exhibited in the envelope and instantaneous phase of the vibration data. The first example involves the analysis of vibration data from a spur gear. The data came from a simple test rig comprising an input gear with 24 teeth driven by an electric motor and meshing with 16 teeth of a pinion¹⁵. The rotational frequencies of the wheel and the pinion are 25 and 37.5 Hz, respectively. This results in the meshing frequency of 600 Hz. The analysed fault is the loss of part of the tooth due to breakage at a point of the working tip. Simply, 1 mm of the facewidth was completely removed. Fig. 2 shows an example of power spectra representing normal and damage conditions. The damage condition spectrum displays a clear pattern of sidebands around the meshing harmonics. Examples of signal demodulation results are given in Fig. 3 and Fig. 4. Time domain averaged signal for damage and normal meshing vibration are given in Fig. 3a and Fig. 4a, respectively. Local tooth fault exhibits an impulse in the envelope function, as shown in Fig. 3b. Also, the change of phase can be observed in the instantaneous phase function in Fig. 3c. In contrast, the instantaneous

characteristics for the normal meshing vibration remain smooth and do not display any disturbances, as shown in Fig. 4b and Fig. 4c.

5.2. Identification of Nonlinear Systems

Fig. 5 shows the impulse response function and its spectrum for the vibration response from a single degree of freedom rig nonlinear test rig with a cubic stiffness characteristic⁴⁶.

The amplitude of the wavelet transform is given in Fig. 6. The ridge of the wavelet transform was computed to obtain the envelope and instantaneous frequency functions for the vibration response. Finally, the backbone curve of the analysed system was constructed to give the result shown by the solid line in Fig. 7. This characteristic clearly displays cubic stiffness nonlinearity. The dashed line in Fig. 7 gives the backbone curve calculated using the Hilbert transform approach. Although the oscillations, which are due to differentiation procedure, can be removed, the results clearly show the advantage of the wavelet analysis over the classical Hilbert transform approach.

6. CONCLUSIONS

The envelope and instantaneous frequency/phase functions, are often used for damage detection in rotating machinery, for identification of nonlinear systems and in deterministic approach to stationarity.

A number of algorithms for the amplitude and frequency/phase demodulation have been briefly described. Finally we are in position to make a comparative analysis. The points to be compared are:

- initial filtering - affects the narrow-band characteristics of the signal;

- determination of the carrier frequency - important since the carrier frequency is not always known or determined with sufficient accuracy;
- integration or differentiation - both operations require a sufficiently high sampling frequency in order to obtain the required accuracy;
- initial amplitude limitation - difficult in numerical implementation.

The comparative results are presented in Table 1. The Hilbert transform and energy distribution algorithms seem to be very attractive. The Hilbert transform method is well established and easy to implement using the classical FFT algorithm. However, it is only valid for single component signals. Also, it requires numerical differentiation for the instantaneous frequency, which is not an easy task. New developments in the area of wavelet analysis can overtake these drawbacks but often lead to expensive computations. Altogether, it appears that the application and implementation very much depend on the problem under consideration.

The envelope and instantaneous frequency/phase functions, are often used for damage detection in rotating machinery, for identification of nonlinear systems and in deterministic approach to stationarity. The paper also shows that the envelope and instantaneous frequency functions are an important link between classical Fourier approach based methods and time-variant analysis.

REFERENCES

- ¹ Randall, R. B. A new method of modeling gear faults, *Transactions of ASME J. Mech. Design*, **104**, 259-267, (1982).
- ² Cempel, C. Diagnostically oriented measures of VA processes, *Journal of Sound and Vibration*, **73**, 547-564, (1983).
- ³ Cempel, C. *Vibroacoustic Condition Monitoring*, Ellis Horwood, Chichester, (1991).
- ⁴ Thrane, N. Zoom – FFT, *Technical Review*, Bruel & Kjaer, 80(2), (1980).
- ⁵ Randall, R. B. *Cepstrum analysis and gearbox fault diagnosis*, Bruel and Kjaer application notes, Naerum, Denmark, (1980).
- ⁶ McFadden, P.D. Determining the location of a fatigue crack in gear from the phase of the change in the meshing vibration, *Mechanical Systems and Signal Processing*, **2**(4), 403-409, (1988).
- ⁷ Mc Fadden, P. D. Detecting fatigue cracks in gears by amplitude and phase demodulation of the meshing vibration, *Transactions of the ASME, Journal of Vib. Acoustics, Stress and Rel. in Design*, **108**, 165-170, (1986).
- ⁸ Mc Fadden, P. D. Low frequency vibration generated by gearbox tooth impacts, *NDT International*, **18**(5), 279-282, (1985).
- ⁹ Cempel, C., and Staszewski, W. J. Use of the signal demodulation, *Zagadnienia Eksploatacji Maszyn*, **84**(4), 505-517, (1990).
- ¹⁰ Cempel, C., Staszewski, W. J. Signal demodulation techniques in vibroacoustical diagnostics of machinery, *Machine Dynamics Problems*, **5**(3), 161-173, (1992).
- ¹¹ Bell, D. *An enveloping technique for detection and diagnosis of incipient faults in rolling element bearings*, Bruel and Kjaer Publication, Naerum, Denmark, (1984).
- ¹² Frarey, J. L. *Bearing surface faults detection*, Mechanical Technology Inc. Publication. Latham, (1987).
- ¹³ Babkin, A. S., and Anderson, J. J. *Mechanical signature analysis of ball bearings*, Federal Scientific Application Note, (1972).

- ¹⁴ Prashad, H., Gosh, M., and Biswas, S. Diagnostic monitoring of rolling-element bearings by high-frequency resonance technique, *Transaction of the American Lubrication Engineers*, 28, 439-448, (1985).
- ¹⁵ Staszewski, W. J. The application of time-variant analysis to gearbox fault detection. PhD Thesis, University of Manchester, Department of Engineering, (1994).
- ¹⁶ Randall, R. B. *Frequency Analysis*, Bruel & Kjaer Application Note, Naerum, Denmark, (1987).
- ¹⁷ Boashah, B., and Jones, G. Instantaneous frequency and time-frequency distributions, *Time-Frequency Signal Analysis*, Ed. B. Boashash, Longman Cheshire, Melbourne, (1992).
- ¹⁸ Tchamitchian, PH., Torresani, B. Ridge and skeleton extraction from the wavelet transform, *Wavelets and their applications*. Ed. M. B. Ruskai, Jones and Bartlett Publishers, Boston, (1992).
- ¹⁹ Cempel, C., *Fundamentals of Vibroacoustical Diagnostics of Machinery*, Warsaw, WNT-Press (In Polish), (1982).
- ²⁰ Lyon, R.H. *Machinery Noise and Diagnostics*, Boston, Butterworth Publishers, (1987).
- ²¹ Staszewski, W., Digital demodulation of signals as a method of vibroacoustical diagnostics. M.Sc. Thesis, Poznan University of Technology (In Polish), (1986).
- ²² McFadden, P.D., and Smith, J. D. Model for the vibration produced by a single point defect in a rolling element bearing, *Journal of Sound and Vibration*, 96(1), 69-82, (1984).
- ²³ McFadden, P. D., and Smith, J. D. The vibration produced by multiple point defects in a rolling element bearing, *Journal of Sound and Vibration*, 98(2), 263-273, (1985).
- ²⁴ Writ, L. S., An amplitude modulation theory for gear-induced vibration, *Strain Gauge Readings*, V(4), (1962).
- ²⁵ McFadden, P.D., An explanation for the asymmetry of the modulation sidebands about the tooth meshing frequency in epicyclic gear vibration, *Proc. Inst. Mech. Eng.*, 199(C1), 65-70, (1985).

- ²⁶ Carlson, A. B. *Communication Systems*, New York, McGraw-Hill Book Company, Third edition, (1986).
- ²⁷ Hsueh, K.D., and Bielawa, R. L. An efficient algorithm for demodulating narrow-band AM signals: theory and implementation, *Journal of Sound and Vibration*, **137**(2), 267-281, (1990).
- ²⁸ Bracewell, R. N. *The Fourier Transform and its Application*, New York, McGrawHill Comp., 2nd ed., (1978).
- ²⁹ Langley, R. S. On various definition of the envelope of a random process, *Journal of Sound and Vibration*, **105**(3), 503-512, (1986).
- ³⁰ Crandall, S. H. Zero crossing, peaks and other statistical measurements of random responses, *Jou. of Acoust. Soc. Am.*, **35**(11), 1693-1699, (1963).
- ³¹ Crandall, S. H. The envelope of random vibration of a lightly damped nonlinear oscillator. *Zagadnienia Drgan Nieliniowych*, **5**, 120-130, (1964).
- ³² Cempel, C. *Vibroacoustical diagnostics*, Poznan Univ. of Tech. Press (in Polish), (1988).
- ³³ Worden, K. Data processing and experiment design for the restoring force surface method. Part I: Integration and differentiation of measured time data, *Mechanical Systems and Signal Processing*, **4**(4), pp. 295-319, (1990).
- ³⁴ Rice, S. O. Definition of envelope function, *Bell Syst. Tech. J.*, **23**, 1-10, (1944).
- ³⁵ Dugundji, J. Envelopes and Pre-envelopes of real waveforms *IRE Transactions on Information Theory*, **IT-4**, 53-57, March (1958).
- ³⁶ Ville, J. A. Théorie et applications de la notion de signal analytique, *Cables et Transmissions*, **2A**, 61-74, (1948).
- ³⁷ McFadden, P. D. Detecting fatigue cracks in gears by amplitude and phase demodulation of the meshing, *Journal of Vib., Acoust., Stress and Rel. in Des.*, **108**, 165-170, (1986).
- ³⁸ Maliszewski, W. Signal demodulation technique in VA diagnostics. M.Sc. Thesis, Poznan University of Technology (In Polish), (1987).
- ³⁹ Kim, Y.C., Khadra, L., and Powers, E. J. Wave modulation in nonlinear dispersive medium, *Phys. Fluids*, **23**(11), 2250-2257, (1980).

⁴⁰ Bloomfield, P. *Fourier Analysis of Time Series*. New York, Wiley, (1976).

⁴¹ Qian, S. and Chen, D. *Joint Time-Frequency Analysis - Methods and Applications*, Prentice Hall PTR, New Jersey, (1996).

⁴² Kronland-Martinet, R., Morlet, J. and Grossmann, A. Analysis of sound patterns through wavelet transforms, *International Journal of Pattern Recognition and Artificial Intelligence*, 1(2), 273-302, (1987).

⁴³ Carmona, R. A., Hwang W. L. and Torresano, B. Characterization of signals by the ridges of their wavelet transforms, *IEEE Transactions on Signal Processing*, 45, 2586- 2594, (1997).

⁴⁴ Staszewski, W. J., Identification of damping in MDOF systems using time-scale decomposition, *Journal of Sound and Vibration*, 203(2), 283-305, (1997).

⁴⁵ Staszewski, W. J., Identification of non-linear systems using multi-scale ridges and skeletons of the wavelet transform, *Journal of Sound and Vibration*, 214(4), 639-658, (1998).

⁴⁶ Staszewski, W.J. and J. E. Chance Identification of nonlinear systems using wavelets – experimental study, In Proc. of the 15th IMAC, Orlando, Florida, pp. 1012-1016, (1997).

FOOD NO. 016 P. 43

Figure 1.

Amplitude and phase modulation signal and its spectrum characteristics.

Figure 2.

Power spectra for spur gear vibration data: (a) normal condition (b) damaged tooth.

Figure 3.

Signal demodulation characteristics for spur gear vibration data representing normal condition: (a) time domain averaged signal (b) envelope (c) instantaneous phase.

Figure 4.

Signal demodulation characteristics for spur gear vibration data representing damaged tooth: (a) time domain averaged signal (b) envelope (c) instantaneous phase.

Figure 5.

Impulse response function (a) and its spectrum for the analysed nonlinear system.

Figure 6.

Wavelet transform of the impulse response function.

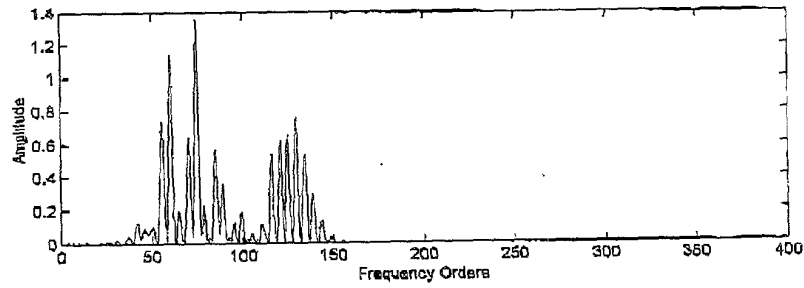
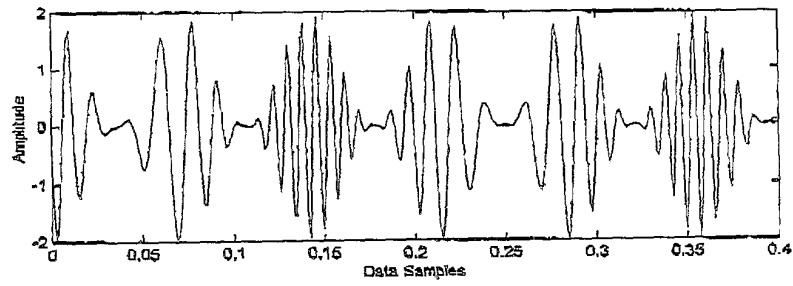
Figure 7.

Backbone curve of the nonlinear system estimated using wavelet analysis (—) and the Hilbert transform (- - -).

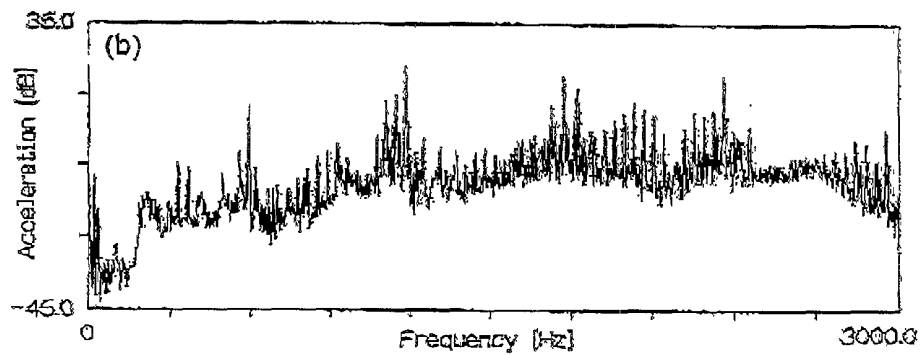
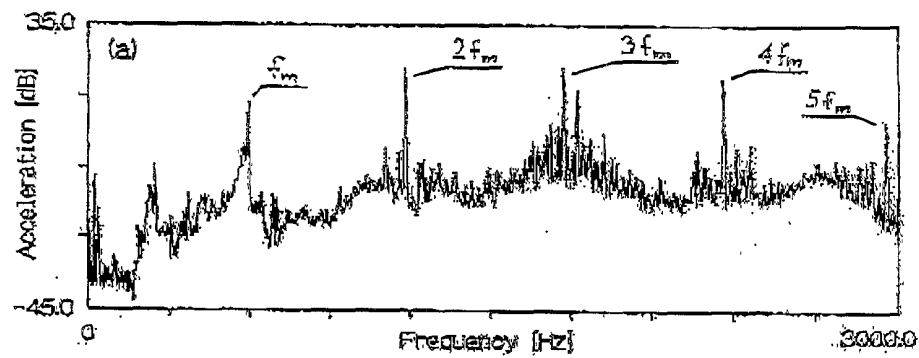
Table 1.

A summary of signal demodulation algorithms.

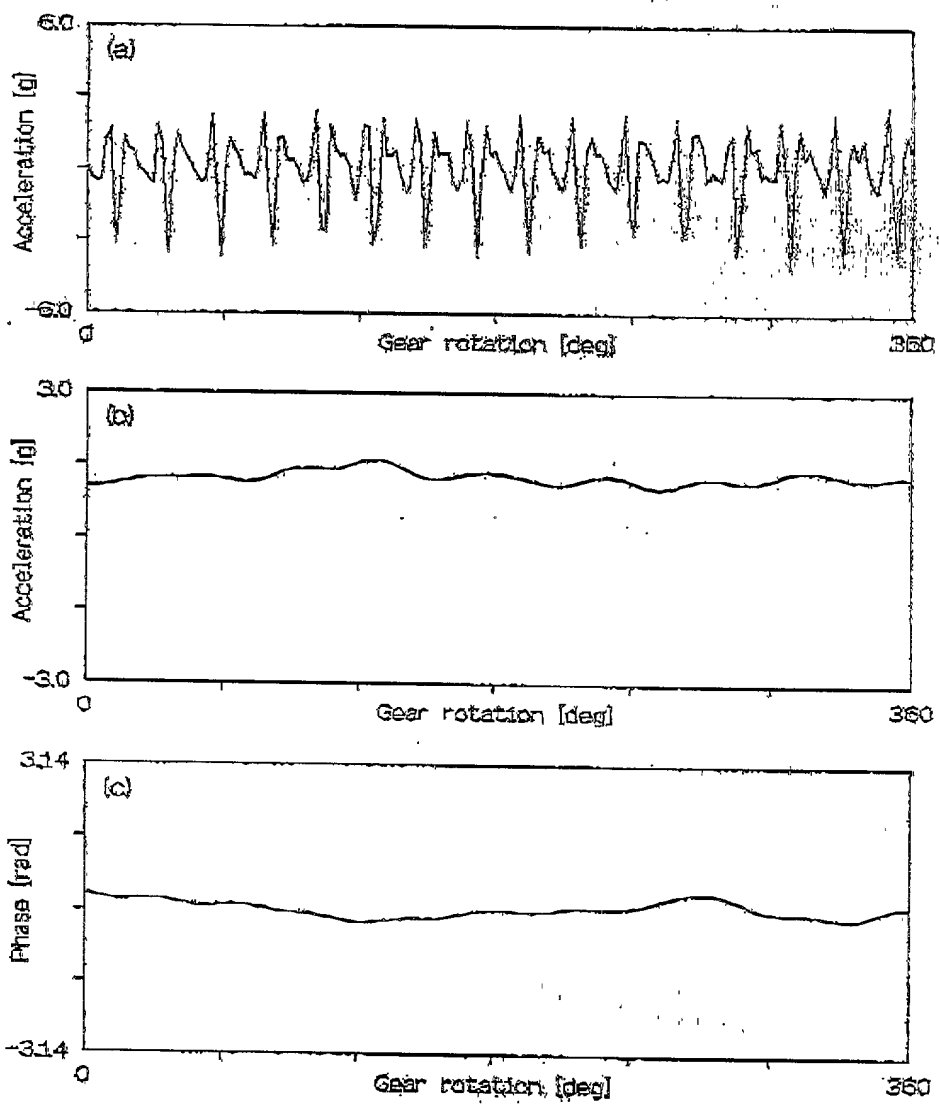
TABLE 1

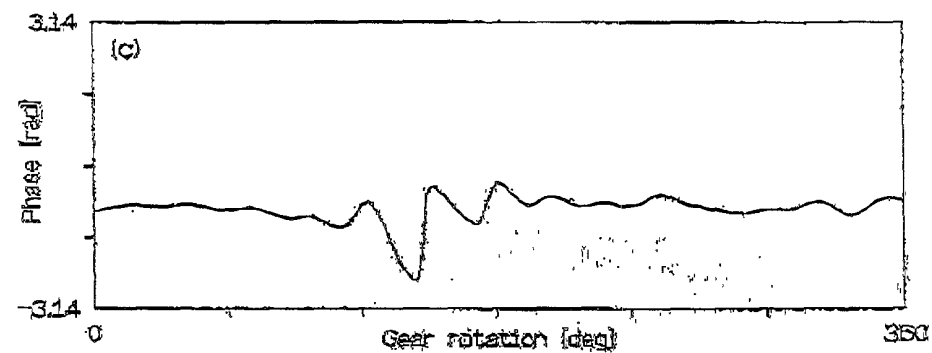
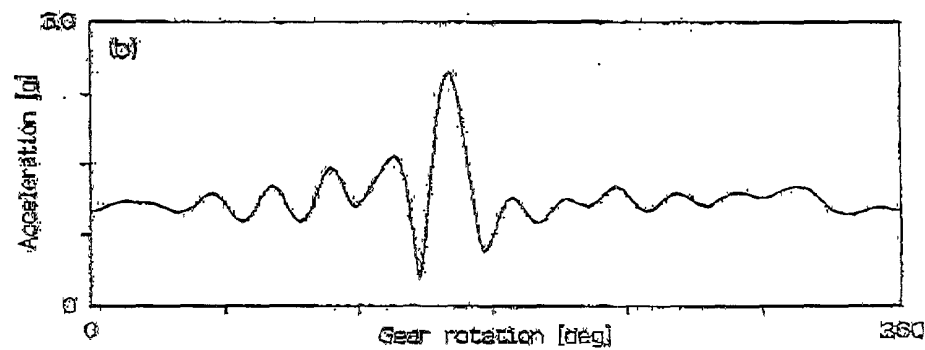
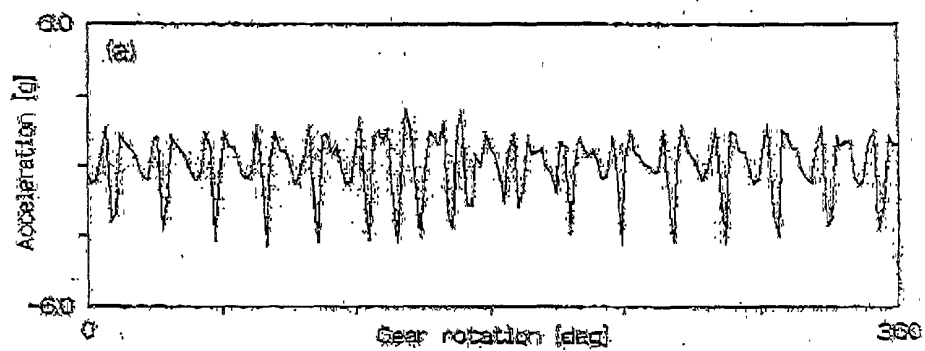


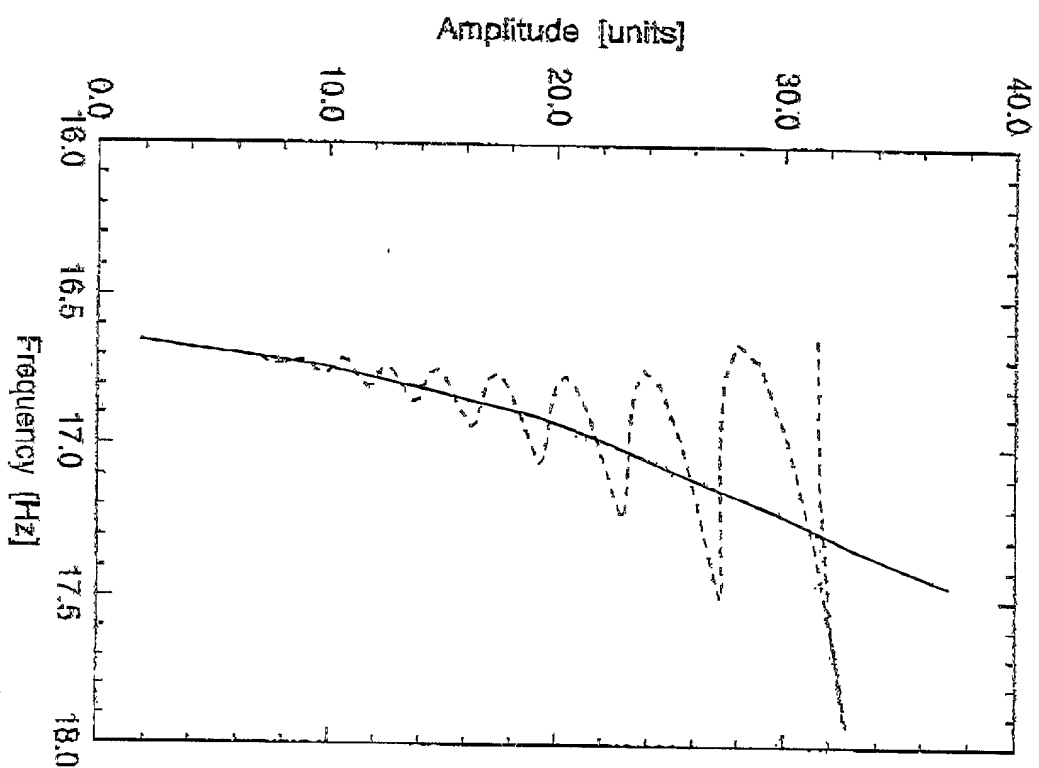
FOODS & DRUGS



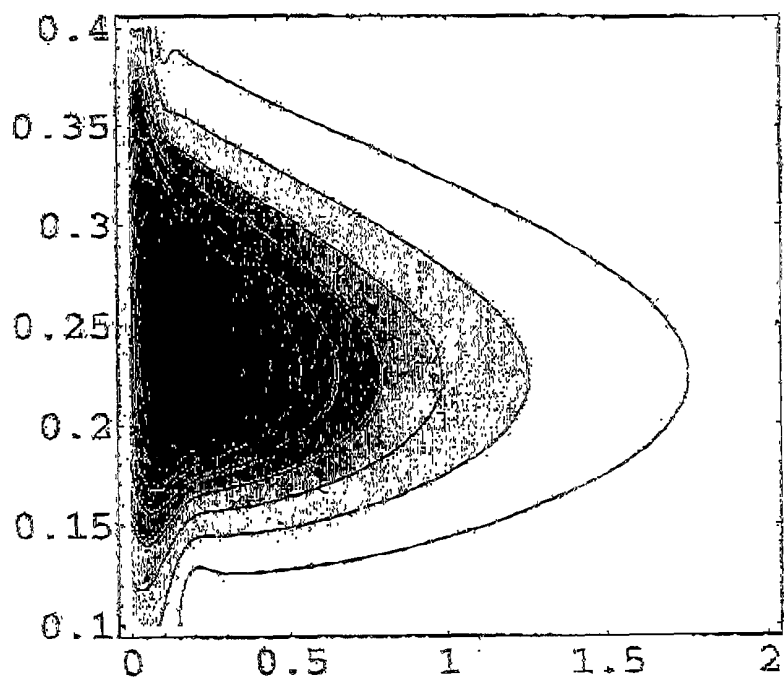
FOUO 2/2/2007



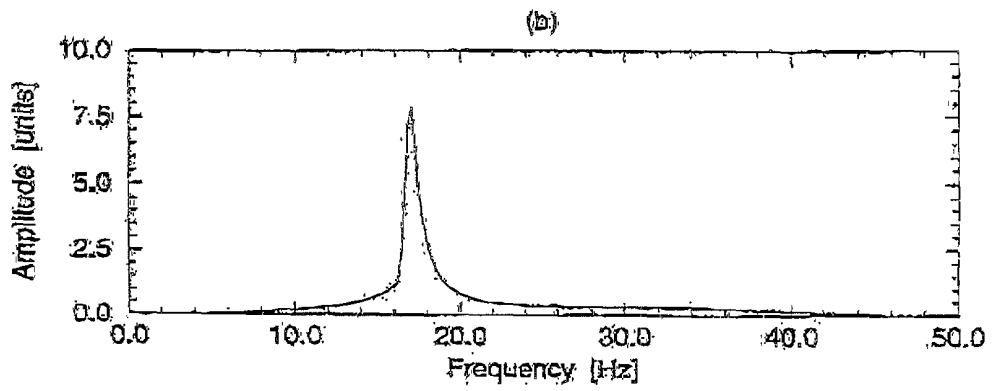
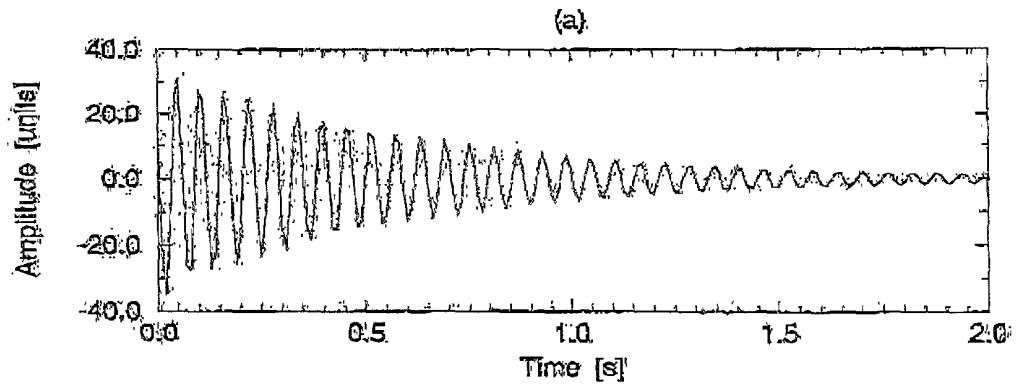




1002773.122001



1002779-10004
FOO22T-2722004



	Carrier frequency required ?	Integration required?	Differentiation required ?	Filtering required ?	Comments
Complex Demodulation Algorithm	Yes	No	No	Yes	-
Local maxima distribution	No	Yes	No	No	Problem with time of averaging
Energy Distribution envelope	No	No	Yes	No	Differentiation can be replaced by integration
Hilbert transform algorithm	No	No	No	No	Well established uses the FFT algorithm but valid only for single component analysis
DFA	Yes	Yes	No	No	-
Wigner-Ville Distribution	No	No	No	Yes	very sensitive to noise in the data
Zero-Crossings	No	No	No	Yes	poor resolution and accuracy
Wavelet Transform	No	No	No	No	suitable for multicomponent signals but computationally expensive